Projection Pursuit via Multivariate Histograms

by

George R. Terrell[1]

Technical Report 85-7, August, 1985

[1]This research was supported by the Office of Naval Research under grant N00014-85-K0100.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
AUG 1985		**00-00-1985 to 00-00-1985**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Projection Pursuit via Multivariate Histograms	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Computational and Applied Mathematics Department ,Rice University,6100 Main Street MS 134,Houston,TX,77005-1892**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**22**	

Abstract

The problem of finding the most interesting low-dimensional subspaces of a multidimensional data set has usually been formulated as a search for the maximum over all projection subspaces of a measure of information. Alternatively, interesting subspaces may be characterized as the eigenspaces associated to the largest eigenvalues of a tensor-valued information measure on the whole space. Since this same information measure solves the problem of the asymptotically optimal mutivariate histogram, the issues of selection and representation are resolved simultaneously. This leads to substantial simplification of both the computational and conceptual problems in projection pursuit.

Key Words: Multivariate Methods; Nonparametric Methods; Density Estimation.

1. Introduction

Statisticians find themselves very often to be taking several measurements simultaneously on each of a number of experimental subjects. An extreme example of this is of course the U. S. census, in which a large number of demographic facts are collected about each of 240,000,000 people. Now if it is reasonable to think of several of these measurements as falling on a continuous real scale, then we are immediately tempted to think of them as coordinates in R^d. This idea goes back at least to Karl Pearson(1901). The tools of affine geometry are then available to study the collection of points that represent the observations.

We find ourselves with two interrelated difficulties with the geometrical interpretation of data. First, analytic geometry assumes homogeneity of the units of measurement in each of the coordinates; this is only fortuitously the case for multivariate statistical data. Two of the coordinates may be (literally) apples and oranges, as in the annual consumption of various fruits by American households. Similarly, if we are to take geometry seriously, there is no mathematical reason to prefer one system of axes to another. Alternative axes may be quite interpretable in the problem under consideration; *apples* $+ 3 \times$ *oranges* might be an interesting index of the contribution of fruits consumed to household nutrition. We may describe the problem mathematically as the requirement to choose the best affine change of variables $\mathbf{y} = A^T\mathbf{x} + \mathbf{b}$ where A is a matrix, $\mathbf{x}$ is the old measurements vector, $\mathbf{y}$ is the new measurements vector, and $\mathbf{b}$ is the new coordinate vector for the origin in the old space.

The second problem is that the statistician may have geometrized the problem in part in order to "see" the problem graphically, in the form of either a scatterplot or a nonparametric density estimate (which may be thought of as a smoothed scatterplot). But these graphics that invoke the ordinary human spatial sense are inherently limited to three dimensions; and except by the use of illusion or changes over time are limited by current graphical technology to two dimensions. Students of graphics are currently experimenting with tricks to overcome this problem; all of them seem quite limited and quickly overburden the human spatial sense as the dimensionality increases. Thus, higher dimensional data challenges the statistician to find low dimensional sub-

spaces of his data space that are of particular interest, especially when represented graphically. It is easy to see that attempts to somehow exhaust possible subspaces are doomed by the combinatorial explosion as the dimension of the original representation rises to quite modest levels. In mathematical notation, we wish to choose a small number of matrices A as in the previous paragraph which may be decomposed $A = A_1|A_2| \cdots |A_k|A_{k+1}$ where A_i, i=1,...,k have at most three columns. These low dimensional coordinate systems are intended to pick out several of the points of view from which our data set is most interesting and interpretable.

This problem is not well-posed unless we are able to provide a mathematical characterization of interesting subspaces. Spearman(1904) proposed that we seek a low dimensional subspace that contained the covariances of the variables, so that the residuals $\mathbf{x} - AA^T\mathbf{x}$ have a diagonal covariance structure. Thus, **factor analysis** declares that a subspace is interesting if several of our measurements are substantially correlated with it. The obvious objection to this democratic criterion is that it is influenced unduly by the choice of measurements to be included in the space. If $\mathbf{x}$ and $\mathbf{y}$ are two independent measurements, then the arbitrary inclusion of the quantity $\mathbf{x}+\mathbf{y}$ as a third variable creates an artificial new factor. Hotelling(1933) suggested that a distinguished coordinate system should consist of the eigenvectors of the covariance matrix of the data set; this is **principal components analysis**. From our point of view, the most interesting subspace would usually be that generated by the eigenvectors associated to the largest eigenvalue. This criterion is unduly influenced by the arbitrary choice of units in any one of our possibly heterogeneous measurements. A change from meters to centimeters in one axis would almost certainly make this a dominant direction by inflating the variance. A standard device for dealing with this problem is to prescale each axis in the same way, say to a variance of one. The effect of this is to emphasize large correlations; and thus to give results similar to factor analysis, with similar difficulties. Thus, these classical methods give very special definitions of important subspaces, which are rather limited in light of the problems we encounter in practice.

Kruskal(1969) suggested that a natural criterion for an interesting subspace was the degree to which the data as seen in that subspace fell readily into more than one cluster. Friedman and Tukey(1974) implemented this approach and called the method **projection pursuit**. Huber(1985)

pointed out that they were maximizing an estimate of the quantity $\int f(x)^2 dx$ where f is the probability density of the data points; he suggested that this was only one possibility for measuring the **information** content of the data as seen from a projection subspace. Huber proposed that more classical measures such as Shannon information $\int f(x) \log f(x) dx$ might be of interest because of certain formal properties they possess. For example, Shannon information attains its minimum over all probability laws with a given covariance matrix at the Gaussian distribution; thus, maximum Shannon information is a criterion that measures nonnormality. Projection pursuit indeed seems to capture a more appealing idea of interest in subspaces than older methods. Marginal densities with high information content show radical asymmetries and multiple modes, which may lead analysts to productive hypotheses about the data set. Nonetheless there are at least two important limitations to these methods. For one thing, their criteria, the measures of information, are somewhat *ad hoc*. A number have been suggested, and though they behave similarly in some ways (though not in others: see Jee(1985)), there is no clear way to choose among them. For another, projection pursuit methods tend to require large amounts of computer time; information criteria tend to be quite complex functions of the choice of subspace (Jee, 1985). Thus, we would like a clearer justification for a particular criterion, and a more computationally tractable method for optimizing it.

This paper will argue that the problems of choosing interesting subspaces of a multivariate data set and of graphical representation of the data are inherently intertwined. For large data sets, particularly in more than two dimensions, the scatterplot is not a very satisfactory data representation. There are problems of masking and of simple overload of one's capacity to interpret so many dots. The usual solution is to use a nonparametric density estimate, which uses smoothing to pick out qualitative features from the sample. The oldest and conceptually simplest nonparametric density estimator is the histogram. Scott(1985) has pointed out that the internal representation of a multivariate histogram, an array of counts, is a particularly tractable form for manipulation of large data sets; and very little information need be lost in the reduction to that form. We will seek an optimal choice of rectangular bin dimensions and grid orientation for a multivariate histogram representation of our data set, by generalizing the approach used in

Scott(1979) for univariate histograms. It will be optimal in the sense that asymptotically the integrated mean squared error $\int E(f - \hat{f})^2$, where f is the true density underlying our data and $\hat{f}$ is a histogram estimate, is to be made least. We will conclude here that the optimal grid orientation for the multivariate histogram is parallel to the principal axes of the quadratic form $\int(\frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j})$; and the optimal bin dimensions are determined by its eigenvalues. This form is a tensor; that is, it transforms in the standard way under changes of variables. It is a natural choice of information measure, somewhat similar in form to the classical Fisher information matrix for multivariate location; we will investigate the conjecture that the subspaces of greatest interest tend to be generated by the eigenvectors corresponding to the largest eigenvalues of the quadratic form, under a suitable preliminary scaling of the data set. Thus, we are identifying the interest in a subspace with the complexity in that direction of the optimal multivariate histogram. Since this matrix is readily estimated from a preliminary histogram of the data, the search for interesting subspaces by this method is very easy computationally. Finally, graphical representation of the subspaces chosen is very readily accomplished from the optimal histogram constructed in the course of the analysis.

2. Multivariate Histograms

A regular rectangular histogram may be constructed by the following device: let $C^d = [-1/2,1/2)\times \cdots \times[-1/2,1/2)$ be the unit cube in R^d centered at the origin. Then every point of R^d falls in a unique subset of the form $\mathbf{z} + C^d$ where $\mathbf{z} \in Z^d$; we will call this the standard bin decomposition of R^d. See Figure 1. If A is a nonsingular square matrix, then an affine bin decomposition of R^d is the set of subsets of the form $\{A\mathbf{x} + \mathbf{o} \mid \mathbf{x} \in \mathbf{C}^d + \mathbf{z}\}$, where $\mathbf{o}$ is called the bin origin. See Figure 2. It turns out that the influence of the choice of bin origin is negligible; and so we will let it be the zero vector in what follows. $\nu_{\mathbf{z}}$ is then the number of elements of a random sample of size n that fall in the bin indexed by $\mathbf{z}$. Then a histogram density estimate is given by

$$\hat{f}(\mathbf{y}) = \frac{\nu_{\mathbf{z}}}{n \mid det\ A \mid}$$

where $\mathbf{y}$ falls in the bin indexed by $\mathbf{z}$.

We wish to choose the matrix A in such a way that asymptotically for n large the integrated mean squared error $\int E(\hat{f}-f)^2$ is least. This error may be decomposed into the integral of the variance and the integral of the squared bias. Without loss of generality we may concentrate on the bin centered at the origin. Assume that the density may be expanded in a Taylor's series about the origin $f(\mathbf{x}) = f(0) + \mathbf{x}^{\mathrm{T}}\nabla \mathrm{f}\,(0) + \mathrm{o}(\mid\mid \mathbf{x} \mid\mid^2)$. The count for that bin is then approximately a Poisson variate with expectation $n \mid det\ A \mid f(0)$. The variance of $\hat{f}$ in that bin is then $f(0)/(\mathrm{n} \mid \det\ \mathrm{A} \mid)$ to first order; and thus the integrated variance is $1/(n \mid det\ A \mid)$ to that order. The bias at a point $\mathbf{y}$ to first order is $\nabla f(0)^{\mathrm{T}}\mathbf{y}$; thus to that order the average bias squared over that bin is

$$\frac{1}{\mid det\ A \mid} \int_{bin} \nabla f(0)^{\mathrm{T}} \mathbf{y}\mathbf{y}^{\mathrm{T}} \nabla \mathrm{f}\,(0) \mathrm{d}\,\mathbf{y}\,.$$

Let us make a change of variables $\mathbf{y} = A\mathbf{x}$; the previous expression becomes

$$\frac{1}{\mid det\ A \mid} \int_{C^d} \nabla f(0)^{\mathrm{T}} \mathrm{A}\,\mathbf{x}\mathbf{x}^{\mathrm{T}} \mathrm{A}^{\mathrm{T}} \nabla \mathrm{f}\,(0) \mathrm{d}\,\mathbf{x}\ \mid \det\ \mathrm{A} \mid$$

$$= \nabla f(0)^{\mathrm{T}} \mathrm{A} \int_{\mathrm{C}^{\mathrm{d}}} \mathbf{x}\mathbf{x}^{\mathrm{T}} \mathrm{d}\,\mathbf{x}\ \mathrm{A}^{\mathrm{T}} \nabla \mathrm{f}\,(0).$$

But $\int \mathbf{x}\mathbf{x}^T$ over the unit cube is just one-twelfth the identity matrix, so we get

$$= \frac{1}{12} \nabla f(0)^{\mathrm{T}} \mathrm{A}\mathrm{A}^{\mathrm{T}} \nabla \mathrm{f}(0).$$

We conclude that the integrated squared bias of the histogram density estimate is to first order

$$\frac{1}{12} \int_{R^d} \nabla f(\mathbf{y})^T A A^T \nabla f(\mathbf{y}) \, d\mathbf{y}$$

$$= \frac{1}{12} \int_{R^d} tr(\nabla f(\mathbf{y})^T A A^T \nabla f(\mathbf{y})) \, d\mathbf{y}$$

and using the trace identity $tr(AB) = tr(BA)$

$$= \frac{1}{12} \int_{R^d} tr(A^T \nabla f(\mathbf{y}) \nabla f(\mathbf{y})^T A) \, d\mathbf{y}$$

$$= \frac{1}{12} tr(A^T \int_{R^d} \nabla f(\mathbf{y}) \nabla f(\mathbf{y})^T \, d\mathbf{y} \, A).$$

As an aside, to simplify our notation let us

Def: For a multivariate density $f(\mathbf{y})$ let the **histogram information matrix**, denoted I_f, be the matrix

$$I_f = \int_{R^d} \nabla f(\mathbf{y}) \nabla f(\mathbf{y})^T d\mathbf{y}.$$

Lemma: When it exists, the histogram information matrix is symmetric, positive definite, and a two-index tensor; that is, given an orthonormal change of variables $\mathbf{x} = O\mathbf{y}$, $I_{g(\mathbf{x})} = O I_{f(\mathbf{y})} O^T$.

Proof: Symmetry is obvious. To see that I_f is positive definite, note that it is the integral over R^d of the nonnegative-matrix-valued function $\nabla f \nabla f^T$ and so is surely nonnegative definite. But if it had a nontrivial null space, the gradient would have to be zero in that direction for almost all $\mathbf{y}$; this is impossible because the density must be positive on a set of positive measure and go to zero along all rays leading out of that set. Thus, the matrix is positive definite. To see that it has tensor character, note that f is a tensor invariant because the Jacobean of an orthonormal change of variables is one. Thus, the matrix of second partials of f is a two-index tensor; further, so is the expectation with respect to f of this matrix. But by integration by parts, $I_f = -E(\partial^2 f / \partial y_i \partial y_j)$.

Q.E.D.

Now we are ready to find the asymptotically optimal bin structure. We wish to minimize

$$IMSE = \frac{1}{n \mid det\ A \mid} + \frac{1}{12} tr(A^T I_f A)$$

by appropriate choice of A. We will find it convenient to reparametrize by letting $h^d = \mid det\ A \mid$ so that h is the edge of a hypercube with the same volume as a bin; and $A = hU$ where U has determinant one. Thus

$$IMSE = \frac{1}{nh^d} + \frac{1}{12} h^2 tr(U^T I_f U)$$

Let us for the moment hold h constant and attempt to minimize the second term with respect to U. The determinant of the matrix of which we are taking the trace is fixed and equal to det I_f. By the convexity of the logarithm, this trace achieves a minimum when the eigenvalues of the matrix with fixed determinant are all equal; that is, when $U^T I_f U$ is a multiple of the identity. There are a multiplicity of solutions U, an issue to which we will return; for the moment we simply notice that the trace we want is $d(det\ I_f)^{\frac{1}{d}}$. Our remaining problem is to minimize

$$IMSE = \frac{1}{nh^d} + \frac{dh^2}{12} (det\ I_f)^{\frac{1}{d}}$$

with respect to h. Differentiating and setting the result equal to 0 we get

$$h = [\frac{6}{n(det\ I_f)^{\frac{1}{d}}}]^{\frac{1}{d+2}}.$$

We conclude

Theorem: An asymptotically optimal histogram density estimate is given by any affine grid formed by multiplying the standard grid on R^d by a matrix A such that

$$* \quad AA^T = (\frac{6}{n})^{\frac{2}{d+2}} (det\ I_f)^{\frac{1}{d+2}}\ I_f^{-1}.$$

This gives

$$IMSE^* = (1+\frac{d}{2}) 6^{\frac{-d}{d+2}} (det\ I_f)^{\frac{1}{d+2}} n^{\frac{-2}{d+2}}.$$

Let us now consider the various solutions A for our optimal histogram. They all give bins of the same volume $\det A = h^d$ and of the same second central moments, by definition. Geometrically, the choice of bins may be thought of as including a variety of shapes of parallelograms subject to the constraints mentioned above. Algebraically, do an eigendecomposition of the right hand side above to get $AA^T = O^T DO$ where O is an orthonormal matrix and D is a diagonal matrix with positive elements. Then the general solution is $A = O^T D^{1/2} P$ where P is any $d \times d$ orthonormal matrix. One very important distinguished special solution is of course the case where the bins are rectangles; this is the usual idea of a multivariate histogram and turns out to have some useful features for the purposes of this paper. A multivariate rectangle is characterized by having mutually perpendicular edges; that is equivalent to $A^T A$ being a diagonal matrix. Since OO^T is the identity, our solution is $A_{rect} = O^T D^{1/2}$; that is, P is the identity. See Figure 3.

I find it interesting that there is one other natural choice of distinguished solution to this problem; as you may remember from high school geometry, a **rhombus** is a parallelogram with equal sides. In general, this would correspond to a choice of A all of whose columns were of the same Euclidean length; equivalently, such that the diagonal elements of $A^T A$ are all equal. We have

Prop: There exists a unique A_{rhomb} which generates the bin structure for an optimal histogram and for which all bin edges are of equal length if I_f has distinct eigenvalues. If this last condition fails, there are multiple rhombic solutions.

The optimal rhomboid histogram may be of interest because its bins exhibit great symmetry and may lead to more aesthetically pleasing statistical graphics. See Figure 4. The concept deserves further study.

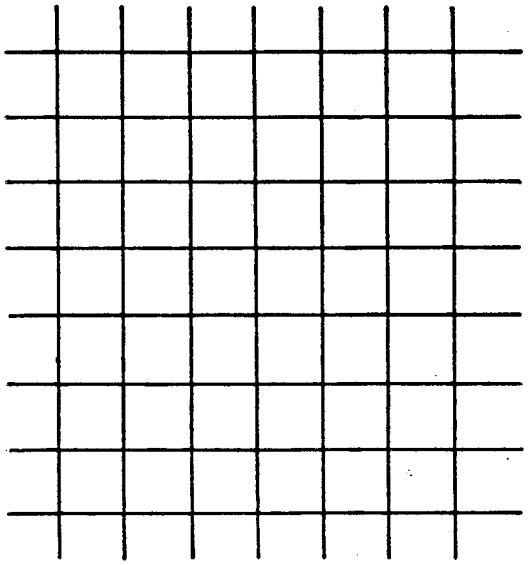

Fig. 1 A Standard grid in $\mathbb{R}^2$

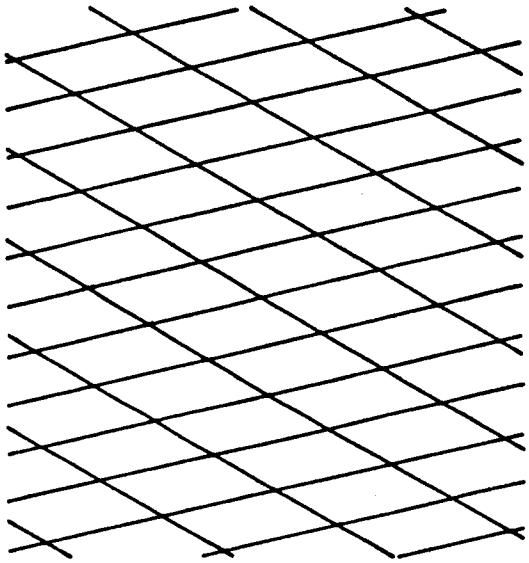

Fig. 2 An Affine grid in $\mathbb{R}^2$

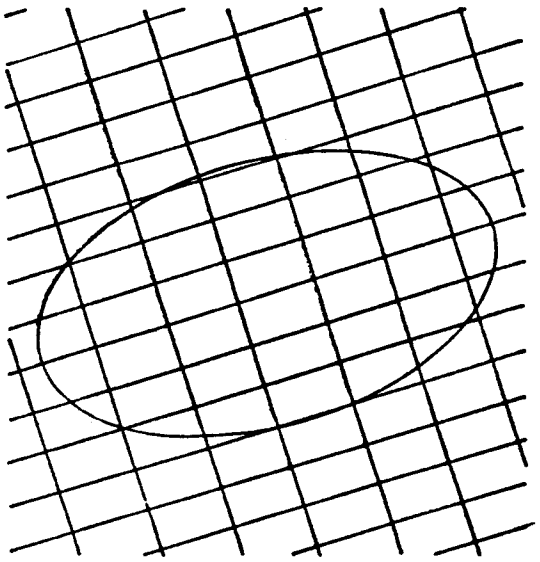

Fig. 3 A Rectangular grid in $\mathbb{R}^2$

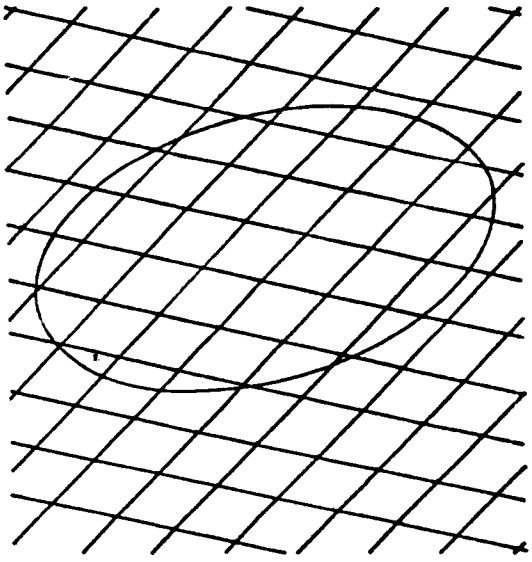

Fig. 4 A Rhomboidal grid in $\mathbb{R}^2$

3. Information Components Analysis

We have suggested that constructing an approximately optimal multivariate histogram may be an aid to choosing low dimensional representations of that data set that are of particular interest. Two questions remain unanswered: how may we construct such a histogram, and precisely how does it direct us to interesting subspaces? The first question may be answered by obtaining an estimate of the histogram information matrix; for that we need a density estimate. It seems reasonable to construct a preliminary histogram; perhaps a maximally smoothed one in the sense of Terrell(1980). A histogram is not differentiable, but the bin counts may plausibly be used to form finite difference estimates of the partial derivatives of the density. It is not difficult to prove

Theorem: If the bin widths of a histogram are of order $n^{\frac{-1}{d+2}}$, then

$$\frac{1}{h_j^2 n^2 h_1 \cdots h_d} \sum_{bins} (\nu_{i_1,\ldots,i_j+1,\ldots,i_d} - \nu_{i_1,\ldots,i_d})^2 - \frac{2}{h_j^2 n h_1 \ldots h_d}$$

is an asymptotically unbiased estimate of $\int (\frac{\partial f}{\partial x_j})^2 = (I_f)_{jj}$; and

$$\frac{1}{h_j h_k n^2 h_1 \ldots h_d} \sum_{bins} (\nu_{i_1} \cdots {}_{i_d} - \nu_{i_1,\ldots,i_j-1,\ldots,i_d})(\nu_{i_1,\ldots,i_d} - \nu_{i_1,\ldots,i_k-1,\ldots,i_d}) - \frac{1}{h_j h_k n h_1 \ldots h_d}$$

is an asymptotically unbiased estimate of $\int (\frac{\partial f}{\partial x_j})(\frac{\partial f}{\partial x_k}) = (I_f)_{jk}$.

I_f, which measures the quality of histogram density estimates, like the other measures of information studies by Huber(1985), measures the bumpiness of the underlying density estimate; since we are particularly interested in especially bumpy densities, it is naturally intertwined with the search for worthwhile graphical representations of the data. Like those measures studied by Huber, it is dependent on the scaling of the original data. If the axes are multiplied by $\tau_1, \ldots, \tau_d$ respectively, then $(I_f)_{j,k}$ is divided by $\tau_j \tau_k \tau_1 \cdots \tau_d$. Since we are more interested in the shape of the density than in scale-dependent features, we shall correct for the original scaling in the course of our analysis. Huber(1985) suggests modification of the information measure; we choose the alternative, equivalent, approach of initially transforming the data to a standard scale. The eigenvectors corresponding to the largest eigenvalues of I_f then indicate the most interesting directions

present in our data. Note a fundamental difference here between study of I_f and projection pursuit; the latter measures the information content of a low dimensional projection, that is, marginal information. By contrast, I_f measures the average bumpiness of conditional densities in each direction; that is, conditional information. Thus in the study of histogram information matrix, which we shall call **information components analysis**, the conditional densities along slices in the principal directions are of as much interest as projections onto that axis.

We propose the following data analysis procedure: in the original multivariate data set, estimate the covariance matrix Σ. Do a principal components analysis (an eigen-decomposition) of this matrix. At this point, we may choose to discard the subspace corresponding to the smallest eigenvalues if the data set seems flat enough in those directions to make them of little interest. Now rescale the data set by multiplication by $\Sigma^{-1/2}$ so that the covariance matrix is the identity. Construct a multivariate histogram by a straightforward generalization of the maximal smoothing principle described in Terrell(1980); we have

Theorem: The d-dimensional density with unit covariance matrix for which the asymptotically optimal histogram has largest bins is

$$f(\mathbf{x}) = (\mathrm{d}+6)^{-\mathrm{d}/2}\frac{\Gamma((\mathrm{d}+6)/2)}{2\pi^{\mathrm{d}/2}}(1-\mathbf{x}^{\mathrm{T}}\mathbf{x}/(\mathrm{d}+6))^2 \quad \text{on} \quad \mathbf{x}^{\mathrm{T}}\mathbf{x} \leq \mathrm{d}+6$$

where the density is zero outside the indicated sphere. Thus the optimal histogram bin for any density with unit covariance matrix has each bin edge

$$h \leq [\frac{3\pi^{d/2}(d+6)}{4n\Gamma((d+6)/2)}]^{1/(d+2)}\sqrt{d+6}.$$

Choosing bin edges to have this maximum size is not only a conservative procedure, it is nearly optimal for the least interesting directions because they have minimum information. Compare Diaconis and Freedman(1982). Now use the histogram to estimate I_f using the Theorem of this section. An eigendecomposition of I_f now allows us to use the Theorem of the previous chapter and formula (*) to construct a second, nearly optimal, histogram aligned along the principal axes of I_f. We are now prepared to study the most interesting subspaces of the data set; projections onto the eigenvectors associated to the largest eigenvalues, and conditional slices in those directions. Notice that all the computations required by this procedure are very easy; all

representations may be constructed by following columns of histogram bins perpendicular to the principal directions.

Thus, we have here proposed a method of exploring many-variable data sets that is as plausible as projection pursuit, somewhat more conceptually coherent, and enormously easier computationally.

4. Some Examples

A pseudorandom sample of size $n=3200$ was drawn from a distribution in which 1100 observations came from a trivariate normal distribution with mean $(-3.0,1.5,1.5)^T$ and unit covariance matrix, 1100 came from one with mean $(1.5,-3.0,1.5)^T$ and unit covariance matrix, and 1000 came from one with mean $(1.5,1.5,-3.0)^T$ and unit covariance matrix. The observations were generated using IMSL routine GGNML. This distribution has well-separated modes which any reasonable density estimate should distinguish easily. The means lie in the plane $x+y+z=0$. This should clearly be the interesting subspace that we would hope would be found by projection pursuit and its relatives.

Our algorithm estimates the covariance matrix of the sample, then transforms it linearly so that its mean is zero and its covariance matrix is the identity. A histogram with cubical bins of edge $h=.75$ (as required by the maximal smoothing principle) is then constructed, and used to estimate I_f by use of the finite difference formula of the last section. IMSL routine EIGRS is then used to extract the eigenvalues and eigenvectors. I illustrate the results when this procedure was applied with seed 111111111.0d0. The eigenvectors have been reexpressed in terms of the original coordinate system:

$$.0178 \quad (.462,.766,.536)^T$$

$$.0560 \quad (1.875,.178,-2.040)^T$$

$$.1014 \quad (-1.254,2.245,-1.110)^T$$

Note that the smallest eigenvalue is much smaller than the other two, and the eigenvectors associated to the two larger eigenvalues both lie close to the plane $x+y+z=0$. A projection into the plane of these two more interesting directions of the estimated optimal histogram is shown in Figure 5; the three clusters stand out very clearly. A number of replications of this experiment with different seeds leads to closely similar results. It is worth noting, and seems to be usually true for samples drawn from this distribution, that the three modes are separated along the second most informative axis; compare Jee(1985).

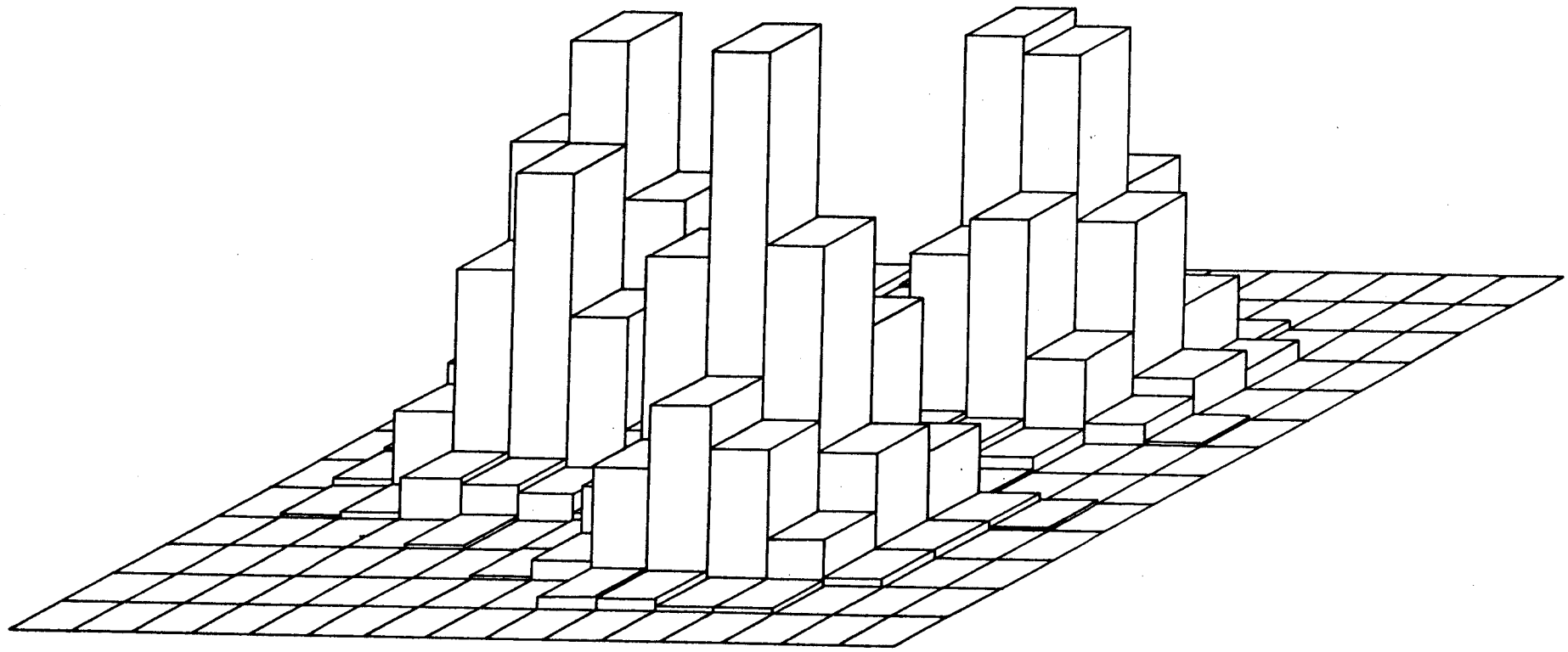

Fig. 5 Best 2-dimensional histogram of a trivariate, trimodal sample (n = 3200)

As a second experiment, 3200 pseudorandom points were generated on the circle $(4cos\,\theta, 4\sqrt{2}/2 sin\,\theta, -4\sqrt{2}/2 sin\,\theta)^T$ (using the IMSL routine GGUBS), and a trivariate normal random vector with unit covariance matrix was added to each. The algorithm was applied to this smoke-ring distribution; with the same seed as in ther example above the eigenvalues and associated eigenvectors were

$$.0105 \quad (-.348, .796, .620)^T$$

$$.0214 \quad (1.175, 1.956, -1.928)^T$$

$$.0328 \quad (-2.717, .746, -.943)^T$$

The two eigenvectors associated to the larger eigenvalues lie close to the plane of the smoke-ring. The histogram projected onto this plane (Figure 6) shows the shape clearly.

Further theory and experiment are needed to discover what sample sizes are required to pin down structures of various degrees of subtlety.

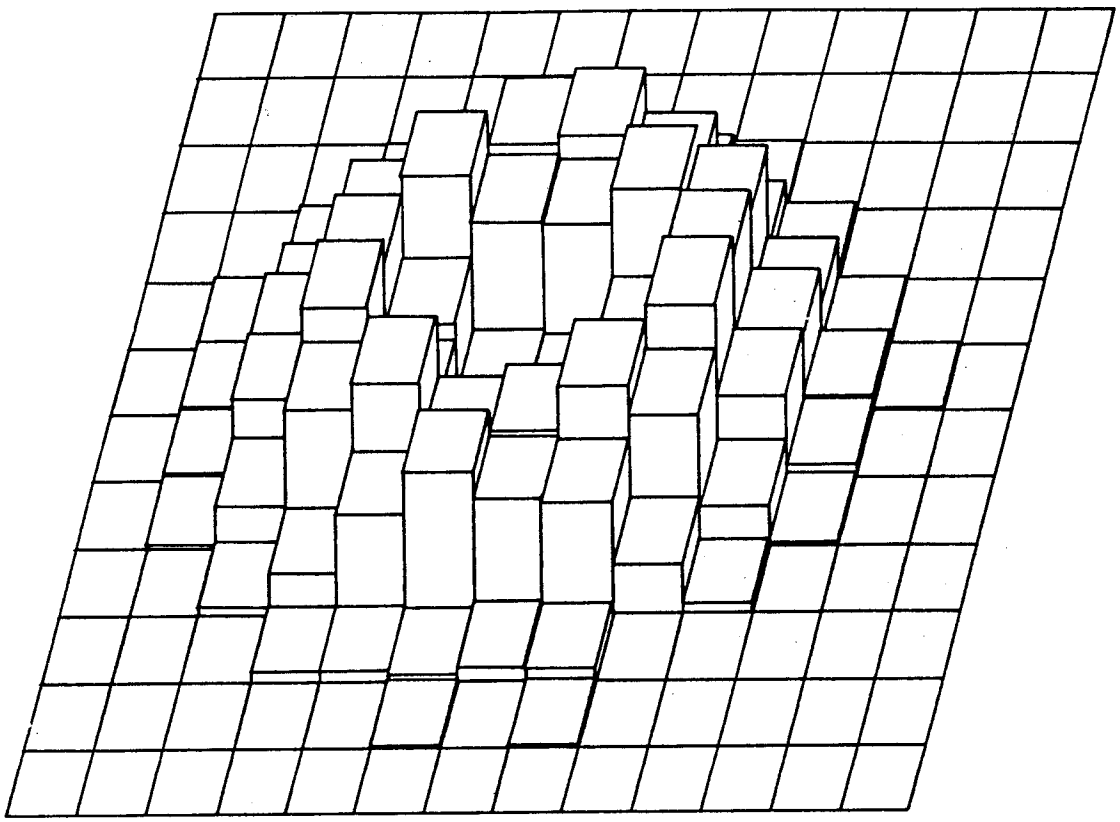

Fig. 6 Best 2-dimensional histogram of a trivariate "smoke ring" sample (n=3200)

5. Acknowledgments

This paper arose out of joint work with David Scott. A series of conversations with Rodney Jee were important in its evolution. Many thanks to both of them.

6. Bibliography

Diaconis, P., and Freedman, D.(1982), "Asymptotics of Graphical Projection Pursuit". *Stanford University Technical Report* Orion 014.

Friedman, J. H., and Tukey, J. W.(1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis". *IEEE Transactions on Computation*, C-23 881-889.

Hotelling, H.(1933), "Analysis of a Complex of Statistical Variables into Principal Components". *Journal of Educational Psychology*, 24 417-441,498-520.

Huber, P. J.(1985), "Projection Pursuit". *The Annals of Statistics*, 13 2 435-474.

Jee, J.R.(1985), "A Study of Projection Pursuit Methods". *Rice University Technical Report*, TR 776-311-4-85.

Kruskal, J. B.(1969), "Toward a Practical Method Which Helps Uncover the Structure of a Set of Multivariate Observations by Finding the Linear Transformation that Optimizes a New Index of Condensation". in *Statistical Computation*, R. C. Milton and J. A. Nelder, editors, Academic Press, New York

Pearson, K.(1901) "On Lines and Planes of Closest Fit to Systems of Points in Space". *Philosophical Magazine*, 6 559-572.

Scott, D. W.(1979) "On Optimal and Data-based Histograms". *Biometrika* 66 605-610.

Scott, D. W.(1985) "Averaged Shifted Histograms: Effective Density Estimators in Several Dimensions". *Annals of Statistics*, to appear.

Spearman, C.(1904) "General Intelligence Objectively Determined and Measured". *American Journal of Psychology* 155 201-293.

Terrell, G. R.(1980) "A Bound for the Smoothing Parameter in Certain Well-known Nonparametric Density Estimators". *Lockheed Technical Memorandum* Lemsco-14850.